DOCUMENT RESUME

ED 307 328 TM 013 486

AUTHOR Batley, Rose-Marie; Boss, Marvin W.
TITLE The Effects on Parameter Estimation of Correlated Dimensions and a Differentiated Ability in a Two-Dimensional, Two-Parameter Item Response Model.
PUB DATE Mar 89
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989). For a related paper, see ED 294 925. Document contains broken print.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Aptitude Tests; Computer Simulation; *Computer Software; *Estimation (Mathematics); Guessing (Tests); *Latent Trait Theory
IDENTIFIERS Differential Ability Tests; Dimensional Analysis; MIRTE Computer Program; *Parametric Analysis; *Two Parameter Model

ABSTRACT

The purpose of this study was to assess the effects of correlated dimensions and differential ability on one dimension on parameter estimation when using a two-dimensional item response theory model. Multidimensional analysis of simulated two-dimensional item response data fitting the M2PL model of M. D. Reckase (1985, 1986) was conducted using the MIRTE analysis program. Six data sets (2,000 ability vectors by 104 items) were generated to satisfy two conditions of the distributions of the ability dimensions and three different degrees of correlation between two abilities. The six data sets (two distributions times three correlations) and analyses were replicated 100 times each. Summary statistics on the 100 replications were used to assess the effects of the degree of correlation between ability dimensions and differential ability on the second dimension. Results indicate that the MIRTE program recovers the structure of a multidimensional correlated space better than do previous estimation programs, especially in the cases in which the items were multidimensional in themselves. However, the MIRTE program tended to underestimate the degree of correlation between the ability dimensions, but it did not force orthogonality on the dimensions. Because of the limitations imposed on any single body of research in terms of research design, some alternative situations need to be studied. Future investigations should assess the accuracy of estimation procedures when a guessing parameter and different latent space structures are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307328

THE EFFECTS ON PARAMETER ESTIMATION OF CORRELATED DIMENSIONS AND A DIFFERENTIATED ABILITY IN A TWO-DIMENSIONAL, TWO-PARAMETER ITEM RESPONSE MODEL

Rose-Marie Batley

Marvin W. Boss

The University of Ottawa

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

ROSE-MARIE BATLEY

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco

March 29, 1989

TM013486

BEST COPY AVAILABLE

Abstract

The purpose of this study was to assess the effects of correlated dimensions and differential ability on one dimension on parameter estimation when using a two-dimensional IRT model. Past research has shown the inadequacies of unidimensional analysis of multidimensional item response data. However, few studies have reported multidimensional analysis of multidimensional data and, in those which used simulated data, results were usually based on one replication.

Multidimensional analysis of simulated two-dimensional item response data fitting the M2PL model of Reckase (1985a, 1985b, 1986) was done using the analysis program, MIRTE (Carlson, 1987).

Six data sets (2000 ability vectors by 104 items) were generated to satisfy two conditions of the distributions of the ability dimensions and three different degrees of correlation between the two abilities. The six data sets (2 distributions x 3 correlations) and analyses were replicated 100 times each. Summary statistics on the 100 replications were used to assess the effects of degree of correlation between ability dimensions and differential ability on the second dimension.

With the exception of the discrimination parameter on the second dimension and the multidimensional discrimination parameter, ability and item parameters were adequately recovered in the data sets in which both abilities were normally distributed over the full range. In the data sets with a restricted range of ability on the second dimension, recovery of the ability and item parameters was adversely affected. As the correlation between the dimensions increased and there was less ability on the second dimension, the dimensions appeared to become less distinguishable. The latent space seemed to be

collapsing into a more unidimensional space when the ability dimensions were correlated 0.50.

Results indicate that MIRTE recovers the structure of a multidimensional correlated space better than previous estimation programs have done, especially in the cases in which the items were multidimensional in themselves. Because of the limitations imposed on any single piece of research in terms of research design, some alternative situations need to be studied. There remains further investigation to be done on the accuracy of estimation procedures when there is inclusion of a guessing parameter as well as with different latent space structures both in terms of population and items.

Theoretical Framework

The original Item Response Theory (IRT) models were based on the assumption of unidimensionality (i.e., only one ability was required to correctly respond to all the items). When more than one ability accounts for test performance, the test is multidimensional and a Multidimensional Item Response Theory (MIRT) model is required to accurately fit the data.

Consider the situation in which items for a test are designed to measure one ability (e.g., mathematics) but require some amount of a second ability (e.g., verbal) in order to respond correctly. This second, required ability could be more crucial to success for some examinees than others. Students of English as a Second Language (ESL) may have sufficient mathematics ability but lack the required amount of verbal ability in order to make a correct response. This could be described as a situation in which mathematics ability is distributed normally over a full range but verbal ability is distributed normally with a lower mean over a narrower range. It is reasonable to assume the two abilities are correlated to some extent. What happens to ability estimates for the ESL students if a MIRT model is used to fit their responses? How are the ability estimates affected by degree of correlation between the abilities?

Several authors (e.g., Ackerman, 1987, Ansley & Forsyth, 1985; Bogan & Yen; 1983, Dorans & Kingston, 1985; Drasgow & Parsons, 1983; McCauley & Mendoza, 1985, McKinley & Reckase, 1984, Reckase, 1979, 1985b, Reckase, Carlson, Ackerman, & Spray, 1986) have considered the effects of analyzing known multidimensional data with a unidimensional item response model. The resulting estimates in most cases were not acceptable unless there was clearly one dominant dimension. Ansley and Forsyth (1985) reported that the unidimensional ability estimates were most highly related to the average of the multidimensional abilities. In the hypothetical educational situation described

above, this would be unacceptable if students with high mathematics ability but low verbal ability were penalized in placement or selection procedures. Reckase et al (1986) found that the unidimensional ability estimates established from multidimensional data had different interpretations at different points on the unidimensional ability scale. By and large, the resulting unidimensional estimates from multidimensional data have been difficult to interpret and have not reflected well the original characteristics of the data.

In spite of findings that unidimensional models are not often robust to multidimensionality, few researchers have made use of multidimensional models to analyze multidimensional data. There are good reasons for this. Although MIRT models are being developed and tested, they are more complex than their unidimensional counterparts. Analysis of multidimensional data with multidimensional programs is expensive in terms of computer time. Few multidimensional analysis programs exist and none has undergone exhaustive testing. Only two programs have been readily available (1) TESTFACT (Wilson, Wood, & Gibbons, 1984), and (2) MAXLOG (McKinley & Reckase, 1983b). TESTFACT has been deemed inappropriate by some researchers because it uses a linear factor analytic procedure to describe the non-linear IRT relationship, a particularly contentious procedure with multidimensional data (Ansley, 1984, Lord, 1980; McDonald & Ahlawat, 1974, R. L. McKinley, personal communication, November 13, 1986). MAXLOG was written to provide parameter estimates for uncorrelated abilities. Results of pilot testing of a third multidimensional analysis program, MIRTE (Carlson, 1987), indicate that it estimates item parameters and abilities more efficiently and more accurately than MAXLOG and it can accommodate data from correlated dimensions. The program is designed to analyze data which fit the multidimensional two-

parameter logistic (M2PL) model (McKinley & Reckase, 1983a, Reckase, 1985b, 1986).

In a test requiring two ability dimensions, if a group of examinees had a normal distribution over the full range of the primary ability but a narrower range and lower mean on the secondary ability, how would this affect parameter estimates? McCauley and Mendoza (1985), in a study of identification of item bias, generated data for items which required a secondary ability on which two groups of examinees differed in mean level. However, the data were generated to conform to a specific factor structure and the analysis was done using a unidimensional model. Their results indicated that differential ability affected the estimates of difficulty moreso than discrimination. The results are not generalizable to multidimensional analysis of multidimensional data.

It is unreasonable to assume abilities are uncorrelated for most achievement tests. McKinley and Reckase (1984) considered the effects of analyzing data generated for correlated dimensions using MAXLOG. The ability and item estimates were confounded in the results of the data analysis. However, when the underlying abilities were correlated and a unidimensional analysis was used, again both unidimensional ability and item parameter estimates were affected (McKinley & Reckase, 1984).

Researchers who have used multidimensional analysis (e.g., McKinley, 1983; McKinley & Reckase, 1983a, 1983b, 1984; Muraki & Englehard, 1985) have indicated that a multidimensional model more adequately describes both real and simulated multidimensional data than does a unidimensional model. However, in most cases, the simulation studies have been based on no replications so that stability of estimates is difficult to determine. There is a need to know how consistently these estimates are recovered. The effects of

both correlated abilities and differential secondary ability on parameter estimation need to be evaluated in a comprehensive, systematic manner.

Purpose of the Study

The purpose of this study was to determine the adequacy of multidimensional ability and item parameter estimates using a MIRT analysis. Specifically three questions were to be addressed.

(1) What is the effect of correlated ability dimensions on parameter estimation for a two-parameter, two-dimensional IRT model?

(2) What is the effect of differential ability on the secondary ability on parameter estimation for the same model?

(3) Are the effects of correlated dimensions similar over the two distributions?

Methodology

A Monte Carlo study was chosen to answer the research questions.

Model Description

The data for the study were generated to fit the multidimensional two-parameter logistic (M2PL) model (McKinley & Reckase, 1983a) which was updated by Reckase (1985a, 1985b, 1986). A description of the updated version follows.

The mathematical formula is given by Equation (1)

$$P_{ij} = P(x_{ij} = 1 \mid \underline{a}_i, d_i, \underline{\theta}_j) = \frac{\exp(\underline{a}_i'\underline{\theta}_j + d_i)}{1 + \exp(\underline{a}_i'\underline{\theta}_j + d_i)}, \qquad (1)$$

$(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, N)$

where P_{ij} is the probability of a correct response to item i by examinee j, x_{ij} is the response (1 = correct, 0 = incorrect) of examinee j on item i, $\underline{a}_i$ is a vector of m discrimination parameters, d_i is a parameter representing the difficulty of item i; $\underline{\theta}_j$ is a vector of m ability parameters for individual j; N is the number of examinees; n is the number of items, and m is the number of dimensions.

This model is compensatory in that it allows high proficiency on one dimension to compensate for low proficiency on other dimensions in arriving at a correct response to a test item.

Reckase (1986) defined a multidimensional discrimination parameter for item i to be

$$MDISC_i = \left[\sum_{k=1}^{m} (a_{ik})^2 \right]^{0.5} \qquad (2)$$

This parameter is related to the item characteristic curve on the multidimensional item response surface above the line through the origin of the ability space and to the point of maximum information and is therefore analogous to the unidimensional discrimination parameter (Carlson, 1987).

Reckase (1985b) also defined a multidimensional item difficulty parameter, $MDIF_i$, such that

$$MDIF_i = -d_i \Big/ \left[\sum_{k=1}^{m} (a_{ik})^2 \right]^{0.5} \qquad (3)$$

$$= -d_i \Big/ MDISC_i$$

This parameter represents the distance between the origin of the m-dimensional ability space and the point in the space where the item information

is a maximum. The line joining this point to the origin is at an angle of α_{ik} to the k^{th} ability dimension where

$$\cos \alpha_{ik} = a_{ik} \Big/ \left[\sum_{k=1}^{m} (a_{ik})^2 \right]^{0.5} \qquad (4)$$

Program Description

The program used to analyze the two-dimensional data was MIRTE (Carlson, 1987). While a version now exists to provide estimates of item and ability parameters for a M3PL model, the version of the program used estimated parameters for the M2PL model. As well as estimation of abilities, item discriminations, and item difficulty, MIRTE provides estimates of standard errors for each of these parameter estimates. Estimates of the multidimensional item difficulty and discrimination are also provided. The method of estimation used is a variation of the joint maximum likelihood procedure using a modified Newton-Raphson iteration technique and the algorithm used is similar to that used in the unidimensional analysis program, LOGIST (Wingersky, Barton, & Lord, 1982). The MIRTE (version 2.00) used in this study was found to estimate parameters when dimensions were correlated better than MAXLOG (J. E. Carlson, personal communication, December, 1987). While MIRTE has been used in one recent study (Ackerman, 1987) to estimate item parameters, the author did not investigate questions considered in this study.

Data Description

Six different data sets were used. The first three sets (A1, A2, A3) represented cases in which both underlying abilities (θ_1 and θ_2) were normally distributed with mean 0, standard deviation 1. The difference among the three sets was the degree of correlation between the abilities, namely 0.00, 0.25, and

0.50. In the second group of data sets (B1, B2, B3) the first ability was normally distributed (mean 0, standard deviation 1) but the second ability had a lower mean and standard deviation (-1 and 0.67 respectively). Again, there were the same three degrees of correlation between the two abilities.

The simulated test consisted of 104 items, 26 items requiring only the first ability, 52 items requiring predominantly the first ability, and 26 items requiring equal amounts of both abilities. A listing of the item parameters is provided in Table 1. Thirteen values of MDIF (ranging from -3 to +3 at intervals of 0.5) and two values of MDISC (2.00, 1.70) were chosen in order to cover the range of difficulties and to simulate realistic discrimination conditions in which the items were designed to discriminate well on the first ability. To meet the requirement that the items discriminate well on the first ability, four values of the angle, α_{i1}, (0°, 15°, 30°, 45°), were chosen. The discrimination indices, a_1 and a_2 (one for each dimension), were then generated to fit the corresponding d and MDISC. The correlations between the original item parameters were: $\rho(d,a_1)$ = 0.004; $\rho(d,a_2)$ = -0.004, $\rho(a_1,a_2)$ = -0.738; and ρ(MDIF,MDISC) = -0.002. Because of the dependency of a_1 and a_2, there is a larger correlation between these parameters. The same item parameters were used for each of the six data sets.

Procedure

The FORTRAN program M2PLGEN (Ackerman, 1985) was used to generate 2000 ability vectors (θ_1,θ_2) satisfying the distributions of θ_1 and θ_2 for Data Set A1. M2PLGEN uses a random seed and the IMSL (1979) subroutine GGNSM to generate random abilities. These ability vectors and the item parameters (a_1, a_2, d) were then used to generate response vectors (0s and 1s) for each of the 2000 simulees to each of the 104 items according to the M2PL model.

Table 1. True Item Parameters for the 104 Items

α_{i1}	$MDIF_i$	d_i	Item	MDISC	a_{i1}	a_{i2}	Item	MDISC	a_{i1}	a_{i2}
0°	3.0	-6	1	2.00	2.00	0.00	53	1.70	1.70	0.00
0°	2.5	-5	2	2.00	2.00	0.00	54	1.70	1.70	0.00
0°	2.0	-4	3	2.00	2.00	0.00	55	1.70	1.70	0.00
0°	1.5	-3	4	2.00	2.00	0.00	56	1.70	1.70	0.00
0°	1.0	-2	5	2.00	2.00	0.00	57	1.70	1.70	0.00
0°	0.5	-1	6	2.00	2.00	0.00	58	1.70	1.70	0.00
0°	0.0	0	7	2.00	2.00	0.00	59	1.70	1.70	0.00
0°	-0.5	1	8	2.00	2.00	0.00	60	1.70	1.70	0.00
0°	-1.0	2	9	2.00	2.00	0.00	61	1.70	1.70	0.00
0°	-1.5	3	10	2.00	2.00	0.00	62	1.70	1.70	0.00
0°	-2.0	4	11	2.00	2.00	0.00	63	1.70	1.70	0.00
0°	-2.5	5	12	2.00	2.00	0.00	64	1.70	1.70	0.00
0°	-3.0	6	13	2.00	2.00	0.00	65	1.70	1.70	0.00
15°	3.0	-6	14	2.00	1.932	0.518	66	1.70	1.642	0.44
15°	2.5	-5	15	2.00	1.932	0.518	67	1.70	1.642	0.44
15°	2.0	-4	16	2.00	1.932	0.518	68	1.70	1.642	0.44
15°	1.5	-3	17	2.00	1.932	0.518	69	1.70	1.642	0.44
15°	1.0	-2	18	2.00	1.932	0.518	70	1.70	1.642	0.44
15°	0.5	-1	19	2.00	1.932	0.518	71	1.70	1.642	0.44
15°	0.0	0	20	2.00	1.932	0.518	72	1.70	1.642	0.44
15°	-0.5	1	21	2.00	1.932	0.518	73	1.70	1.642	0.44
15°	-1.0	2	22	2.00	1.932	0.518	74	1.70	1.642	0.44
15°	-1.5	3	23	2.00	1.932	0.518	75	1.70	1.642	0.44
15°	-2.0	4	24	2.00	1.932	0.518	76	1.70	1.642	0.44
15°	-2.5	5	25	2.00	1.932	0.518	77	1.70	1.642	0.44
15°	-3.0	6	26	2.00	1.932	0.518	78	1.70	1.642	0.44
30°	3.0	-6	27	2.00	1.732	1.00	79	1.70	1.472	0.85
30°	2.5	-5	28	2.00	1.732	1.00	80	1.70	1.472	0.85
30°	2.0	-4	29	2.00	1.732	1.00	81	1.70	1.472	0.85
30°	1.5	-3	30	2.00	1.732	1.00	82	1.70	1.472	0.85
30°	1.0	-2	31	2.00	1.732	1.00	83	1.70	1.472	0.85
30°	0.5	-1	32	2.00	1.732	1.00	84	1.70	1.472	0.85
30°	0.0	0	33	2.00	1.732	1.00	85	1.70	1.472	0.85
30°	-0.5	1	34	2.00	1.732	1.00	86	1.70	1.472	0.85
30°	-1.0	2	35	2.00	1.732	1.00	87	1.70	1.472	0.85
30°	-1.5	3	36	2.00	1.732	1.00	88	1.70	1.472	0.85
30°	-2.0	4	37	2.00	1.732	1.00	89	1.70	1.472	0.85
30°	-2.5	5	38	2.00	1.732	1.00	90	1.70	1.472	0.85
30°	-3.0	6	39	2.00	1.732	1.00	91	1.70	1.472	0.85
45°	3.0	-6	40	2.00	1.414	1.414	92	1.70	1.202	1.202
45°	2.5	-5	41	2.00	1.414	1.414	93	1.70	1.202	1.202
45°	2.0	-4	42	2.00	1.414	1.414	94	1.70	1.202	1.202
45°	1.5	-3	43	2.00	1.414	1.414	95	1.70	1.202	1.202
45°	1.0	-2	44	2.00	1.414	1.414	96	1.70	1.202	1.202
45°	0.5	-1	45	2.00	1.414	1.414	97	1.70	1.202	1.202
45°	0.0	0	46	2.00	1.414	1.414	98	1.70	1.202	1.202
45°	-0.5	1	47	2.00	1.414	1.414	99	1.70	1.202	1.202

Table 1. (cont.) True Item Parameters for the 104 Items

α_{i1}	$MDIF_i$	d_i	Item	MDISC	a_{i1}	a_{i2}	Item	MDISC	a_{i1}	a_{i2}
45°	-1.0	2	48	2.00	1.414	1.414	100	1.70	1.202	1.202
45°	-1.5	3	49	2.00	1.414	1.414	101	1.70	1.202	1.202
45°	-2.0	4	50	2.00	1.414	1.414	102	1.70	1.202	1.202
45°	-2.5	5	51	2.00	1.414	1.414	103	1.70	1.202	1.202
45°	-3.0	6	52	2.00	1.414	1.414	104	1.70	1.202	1.202

The 2000 x 104 matrix of response vectors was analyzed using MIRTE to provide estimates of θ_1, θ_2, a_1, a_2, d, MDIF, MDISC, α_1, and α_2. These results were filed, the random seed was incremented by two and the process was repeated. For Data Set A1 there were 100 replications. Summary statistics were calculated on the 100 replications.

This procedure was repeated for the other five data set conditions. The same initial item parameter estimates for a_1 and a_2 were used for every replication in order to provide better control in the design. Finally, summary results from the six data sets were compared.

Each job of 100 replications required approximately 45,000 to 50,000 CPU seconds. The jobs were run in batch on an Amdahl 5880 processor with 64 megabytes of main memory. The VM/HPO operating system was in use.

Results and Discussion

The purpose of this research was to determine the effects of correlated abilities and differential ability on one dimension on parameter estimation given a two-dimensional, two-parameter logistic item response model. First it should be determined if suitable ability data were generated to model the conditions specified. Then it needs to be determined whether MIRTE adequately estimated the parameters from the analysis of the response vectors generated. Results are

discussed in Part 1 for Data Sets A1, A2 and A3, in Part 2 for Data Sets B1, B2 and B3 and in Part 3 for comparisons made among the A and B data sets. The statistics given in this section are the mean values of the corresponding statistics determined for each of the 100 replications in each data set.

Part 1

Generation of (θ_1, θ_2). The ability data in all three data sets were generated to fit the specifications stated. The correlation between θ_1 and θ_2 for data generated over the 100 replications was recovered as -0.001 for Data Set A1, 0.251 for A2, and 0.500 for A3. The means for θ_1 and θ_2 were in the range 0.002 to -0.004 and standard deviations were within 1± 0.003. There was very small variance (less than 0.0005) for these means and standard deviations in all data sets. There were no replications in which the ability data were not satisfactorily generated.

In keeping with the findings of Greaud (1988), the mean raw score appeared to be unaffected by changes in degree of correlation between the ability dimensions. (All raw score means were approximately 52.)

Recovery of Ability Parameters. In each of the three data sets over the 100 replications, $\hat{\theta}_1$ and $\hat{\theta}_2$ had means of 0.00 and standard deviations of 1.00. The standard deviation of the mean was less than 0.001 for all data sets. The recovery of these statistics is not particularly meaningful as a measure of accuracy in these cases because the MIRTE program rescales the theta estimates to mean 0, standard deviation 1 after each iteration in order to prevent drifting of the estimates.

In the data analysis, the program doesn't always identify dimensions one and two correctly. In order to avoid confusing the dimensions during the 100 replications, a check was made during each data analysis on the first thirteen item discrimination parameter estimates. (These items were pure on θ_1.) If the

sum of the first thirteen a_1 estimates was less than the sum of the first thirteen a_2 estimates, the estimations for the dimensions were flipped.

The mean average absolute deviation of $\hat{\theta}_1$ from the true θ_1 ($AAD(\hat{\theta}_1)$) ranged from 0.446 to 0.459 (see Table 2). (Note that the tables appearing in the text contain results for all six data sets in order to save space and so that comparisons can be seen more readily.) Increasing $\rho(\theta_1,\theta_2)$ did not appear to affect this. The mean average absolute deviation of $\hat{\theta}_2$ ($AAD(\hat{\theta}_2)$) ranged from 0.544 to 0.412 and seemed to be more affected by the correlation between the abilities. As $\rho(\theta_1,\theta_2)$ increased, the $AAD(\hat{\theta}_2)$ decreased. This is probably because of the compensatory nature of the M2PL model. There was very little variance over replications in these AADs (0.001 for $\hat{\theta}_1$, 0.002 for $\hat{\theta}_2$) so that the thetas appear to have been recovered consistently across the three data sets.

Table 2. Mean Values of Statistics for Estimated Thetas (over 100 replications)

Data Set	$\rho(\theta_1,\theta_2)$	$AAD(\hat{\theta}_1)$	$AAD(\hat{\theta}_2)$	$r(\hat{\theta}_1,\hat{\theta}_2)$	$r(\theta_1,\hat{\theta}_1)$	$r(\theta_2,\hat{\theta}_2)$	$r(\theta_1,\hat{\theta}_2)$	$r(\theta_2,\hat{\theta}_1)$
A1	0.00	0.447	0.544	0.062	0.842	0.764	0.505	-0.295
A2	0.25	0.446	0.470	0.179	0.842	0.824	0.603	-0.050
A3	0.50	0.459	0.412	0.282	0.831	0.865	0.699	0.209
B1	0.00	0.463	0.856	0.147	0.773	0.517	0.662	-0.170
B2	0.25	0.544	1.079	0.201	0.765	0.623	0.713	0.052
B3	0.50	0.566	1.047	0.218	0.744	0.721	0.755	0.247

The relationship between the ability parameter θ_1 and its estimate was adequately recovered as $r(\theta_1,\hat{\theta}_1)$ was greater than 0.83 for all three data sets. In Data Set A3, θ_2 appeared to be recovered better than θ_1 in spite of the fact that few items were measuring the θ_2-space. This was also supported by the decreasing $AAD(\hat{\theta}_2)$ as the correlation between the ability dimensions increased. As $\rho(\theta_1,\theta_2)$ increased, θ_1 was less well recovered but θ_2 was better recovered. This was supported by the mean correlation between θ_2 and $\hat{\theta}_2$. As $\rho(\theta_1,\theta_2)$

increased. θ_2 became more highly correlated with $\hat{\theta}_2$ (Table 2). In all three data sets, θ_2 was recovered fairly well according to $r(\theta_2,\hat{\theta}_2)$.

The mean standard error of the thetas (as calculated by MIRTE) was approximately 0.259, almost half the size of the AADs. The variance in these mean standard errors was very small although the standard errors were more spread out as the correlation between the dimensions increased.

The correlation between the ability dimensions was not well recovered. As $\rho(\theta_1,\theta_2)$ increased, MIRTE tended to produce ability estimates which were less correlated than the generated abilities. The difference between $\rho(\theta_1,\theta_2)$ and $r(\hat{\theta}_1,\hat{\theta}_2)$ increased as $\rho(\theta_1,\theta_2)$ increased. This result agrees with that reported by Carlson (1987).

Recovery of Item Parameters. In the maximum likelihood estimation procedures used in MIRTE, ability estimates are used to improve item parameter estimates and vice versa. Hence, the final estimates are affected by each other. As $\rho(\theta_1,\theta_2)$ increased, what happened to the item parameter estimates?

Statistics on the item difficulty parameters are summarized in Table 3. In all three data sets, $r(d,\hat{d}) = 0.997$ indicating good recovery of the relationship between the item difficulty parameter and estimate. As $\rho(\theta_1,\theta_2)$ increased, the mean and standard deviation of $\hat{d}$ were increasingly overestimated but remained close to the original parameter statistics. The $AAD(\hat{d})$ increased slightly as the correlation between the ability dimensions increased indicating that d was being less well recovered. However, the standard error of $\hat{d}$ decreased as $\rho(\theta_1,\theta_2)$ increased. The mean and standard deviation of the multidimensional difficulty parameter, MDIF, were recovered well although here again MDIF was less well recovered as $\rho(\theta_1,\theta_2)$ increased. MDIF is a function of the discrimination parameters and its estimate is therefore affected by the estimates of the a_i parameters.

Table 3. Summary of Mean Statistics for Item Difficulty (over 100 replications)

Data Set	$\hat{d}$	$s(\hat{d})$	$se(\hat{d})$	$AAD(\hat{d})$	$\widehat{MDIF}$	$s(\widehat{MDIF})$	$r(d,\hat{d})$	$r(MDIF,\widehat{MDIF})$
True	0.009	3.771	-----	-----	-0.005	2.058	-----	-----
A1	0.009	3.929	0.112	0.224	0.006	2.079	0.997	0.995
A2	0.010	3.936	0.109	0.228	0.005	2.028	0.997	0.994
A3	0.030	3.936	0.106	0.232	0.012	1.999	0.997	0.991
B1	-0.726	3.995	0.137	0.811	0.460	2.535	0.984	0.958
B2	-0.734	4.001	0.132	0.827	0.434	2.459	0.982	0.956
B3	-0.716	4.044	0.123	0.834	0.397	2.247	0.982	0.969

s - standard deviation; se - standard error from MIRTE program

Discrimination parameter estimates have been reported to be affected more by multidimensional data. This result was also evident in this study. The mean of $\hat{a}_1$ was lower than the true mean and the standard deviation was higher than the true standard deviation for all three data sets (see Table 4). The mean of $\hat{a}_2$ was much higher than the true mean of 0.678. In fact the mean of $\hat{a}_2$ was higher than the mean estimates of a_1 and approached the true mean of a_1 as $\rho(\theta_1,\theta_2)$ increased. Both means increased slightly as $\rho(\theta_1,\theta_2)$ increased. The standard deviation of $\hat{a}_2$ was higher than the true standard deviation but there was not as large a difference here as with $\hat{a}_1$. Standard errors of estimation of $\hat{a}_1$ and $\hat{a}_2$ were approximately 0.09 but the AADs were much larger, particularly for $\hat{a}_2$. As the correlation between the two ability dimensions increased, the $AAD(\hat{a}_2)$ increased slightly indicating a_2 was being less well recovered. The $AAD(\hat{a}_1)$ was approximately 0.5 for all three data sets. The standard errors of both the discrimination parameter estimates were similar in size but $se(\hat{a}_2) \leq se(\hat{a}_1)$.

The multidimensional discrimination parameter, MDISC, was recovered with a higher mean and higher standard deviation in all three data sets. There appears to be a rotational indeterminacy in the recovery of the discrimination

parameters and a tendency to spread the discrimination parameter estimates over the entire space even though they originally did not cover the entire space

Table 4. Summary of Mean Statistics for Item Discrimination (over 100 replications)

Data Set	$\hat{a}_1$	$s(\hat{a}_1)$	$se(\hat{a}_1)$	$AAD(\hat{a}_1)$	$\hat{a}_2$	$s(\hat{a}_2)$	$se(\hat{a}_2)$	$AAD(\hat{a}_2)$	$\widehat{MDISC}$	$s(\widehat{MDISC})$	$\hat{\alpha}_1$
True	1.637	0.251	-----	-----	0.678	0.496	-----	-----	1.850	0.151	22.50
A1	1.195	0.569	0.099	0.500	1.379	0.512	0.096	0.707	1.957	0.288	49.07
A2	1.201	0.528	0.095	0.486	1.448	0.582	0.094	0.775	2.013	0.319	49.40
A3	1.202	0.502	0.093	0.490	1.510	0.628	0.093	0.836	2.057	0.361	49.98
B1	1.076	0.551	0.119	0.623	1.228	0.449	0.112	0.653	1.736	0.398	49.09
B2	1.094	0.557	0.138	0.620	1.298	0.501	0.108	0.708	1.803	0.449	49.76
B3	1.139	0.599	0.134	0.624	1.408	0.534	0.103	0.791	1.922	0.495	50.94

s - standard deviation, se - standard error from MIRTE program

This was supported by the statistics on the angle estimates, α_1 and α_2. Originally α_1 had a mean of 22.50°. This was recovered in all data sets at over 49° Similarly, α_2, whose original mean was 67.50°, was recovered in all data sets at just over 40°. The original standard deviation of 16.85° increased for the estimates to approximately 20° There seemed to be an attempt to cover the entire $\theta_1\theta_2$-space in estimation of parameters related to discrimination. Estimates of α_1 and α_2 ranged from very close to 0° to almost 90°

Correlation coefficients again were used to determine adequacy of the parameter recovery (Table 5). In all cases, a_1 correlated more highly with $\hat{a}_1$ than with $\hat{a}_2$. Similarly, a_2 correlated more highly with $\hat{a}_2$ than it did with $\hat{a}_1$. As well, a_2 correlated higher with $\hat{a}_2$ than a_1 did with $\hat{a}_2$. The anomaly in the correlations was that a_1 correlated less highly with $\hat{a}_1$ than a_2 did with $\hat{a}_1$. As the discrimination parameter estimates appear to be dispersed across the $\theta_1\theta_2$-space, this may account for the apparent better recovery of a_2 than of a_1. That the standard deviation of a_2 was twice as large as that of a_1 may also account for the higher correlations of both $\hat{a}_1$ and $\hat{a}_2$ with a_2 The greater variability in

a_2 would allow for higher correlations. The AAD($\hat{a}_2$) did not support the conclusion that a_2 was better recovered than a_1.

The correlation between $\hat{a}_1$ and $\hat{a}_2$ was slightly stronger than the true parameter correlation of -0.738 except in the Data Set A3 where it was slightly smaller. The multidimensional discrimination parameter, MDISC, did not correlate as highly with its estimate. This correlation was highest (0.600) when the ability dimensions were uncorrelated and decreased as the correlation between the abilities increased.

Table 5. Mean Correlations for Item Discrimination Values (over 100 replications)

Data Set	$r(a_1,\hat{a}_1)$	$r(a_2,\hat{a}_2)$	$r(\hat{a}_1,\hat{a}_2)$	$r(a_1,\hat{a}_2)$	$r(a_2,\hat{a}_1)$	$r(MDISC,\widehat{MDISC})$	$r(\alpha_1,\hat{\alpha}_1)$
A1	0.834	0.893	-0.765	-0.572	-0.865	0.600	0.943
A2	0.818	0.899	-0.769	-0.587	-0.830	0.565	0.933
A3	0.760	0.895	-0.735	-0.587	-0.747	0.502	0.907
B1	0.530	0.523	-0.428	-0.309	-0.543	0.296	0.630
B2	0.460	0.511	-0.401	-0.306	-0.455	0.269	0.586
B3	0.431	0.514	-0.459	-0.285	-0.403	0.285	0.564

Carlson (1987) reported that estimates of the discrimination parameters are sensitive to the distribution of the discrimination parameters in the generated data. The a_1 parameters were not distributed over the entire latent space. This restricted then the recovery of these parameters which, in turn, affected the recovery of the ability and difficulty parameters. That the items of the simulated test did not cover the entire latent space and the variability in a_2 was so much greater than in a_1 would both affect recovery of parameters in a detrimental way.

Part 2

Generation of (θ_1,θ_2): Again the ability data in the three B data sets were generated to fit the specifications stated. The correlation between θ_1 and θ_2 for

data generated over the 100 replications was recovered as -0.001 for Data Set B1, 0.251 for B2, and 0.499 for B3. The means for θ_1 were in the range of -0.002 to -0.004 with a standard deviation range of 1.000 to 1.003, for θ_2 the means were in the range -0.999 to -1.001 with a standard deviation range of 0.669 to 0.671. Again there was very small variance (less than 0.0005) for these means and standard deviations in all data sets. There were no replications in which the ability data were not satisfactorily generated.

The raw score on the test was affected by the differentiated ability on θ_2. The raw score means were about 5 points lower at approximately 47. Increasing the correlation between the ability dimensions did not appear to affect the raw score mean.

Recovery of Ability Parameters. In each of the three data sets over the 100 replications, θ_1 and θ_2 had means of 0.00 and standard deviations of 1.00. As the MIRTE program rescales the theta estimates to mean 0, standard deviation 1 after each iteration, these estimates cannot be meaningfully compared to the means and standard deviations of the generated parameters.

The AAD($\hat{\theta}_1$) ranged from 0.463 to 0.566 (see Table 2 above). As $\rho(\theta_1,\theta_2)$ increased, AAD($\hat{\theta}_1$) increased. The values of AAD($\hat{\theta}_2$) ranged from 0.856 to 1.047 and changes in $\rho(\theta_1,\theta_2)$ didn't produce predictable changes in this statistic. The rescaling of $\hat{\theta}_2$ is reflected in the values of AAD($\hat{\theta}_2$). The variance in these statistics over the replications was very small as in the A data sets. The mean standard error of the theta estimates in the B data sets was 0.287, larger than that in the A data sets.

The recovery of the relationship between the parameter and its estimates was not as high as in the A data sets for either θ_1 or θ_2. The relationship between θ_1 and its estimate was greater than 0.74, between θ_2 and its estimate greater than 0.52 (see Table 2 above). As $\rho(\theta_1,\theta_2)$ increased, the $r(\hat{\theta}_1,\hat{\theta}_2)$ also

increased. Increasing the correlation between the dimensions had the same effect in the B data sets as in the A data sets, i.e., θ_1 appeared to be less well recovered and the recovery of θ_2 improved.

The correlation between the ability dimensions was not well recovered. As in the A data sets, MIRTE produced ability estimates which were less correlated than the generated abilities when $\rho(\theta_1,\theta_2) \neq 0$.

Neither θ_1 nor θ_2 was recovered as well in the B data sets as in the A data sets. It seemed to be more difficult for the program to distinguish between the dimensions and there was a greater tendency to collapse the space.

Recovery of Item Parameters Statistics on the item difficulty parameters are provided in Table 3 (above). Both d and MDIF were less well recovered in the B data sets than in the corresponding A data sets. The rescaling of θ_2 to mean 0, standard deviation 1 in the MIRTE program made the estimates of the theta vectors for the "sample" in the B data sets appear more able than the original theta vectors would indicate. This resulted in the items appearing to be more difficult than they were. There were larger standard errors for $\hat{d}$ and larger standard deviations for $\hat{d}$ and $\widehat{MDIF}$ than in the A data sets. As $\rho(\theta_1,\theta_2)$ increased, $se(\hat{d})$ decreased but $AAD(\hat{d})$ increased. The $AAD(\hat{d})$ were much larger than in the A data sets. The mean of MDIF was greatly overestimated. However, the recovery of the relationship between the parameter and its estimate remained high. The correlation $r(d,\hat{d}) > 0.98$ and $r(MDIF,\widehat{MDIF}) > 0.95$ and there was little change in these correlations as $\rho(\theta_1,\theta_2)$ increased.

The estimates of the discrimination parameters were similar to those for the A data sets (Table 4 above). The parameter a_1 was underestimated, a_2 was overestimated, and the mean estimate of a_2 was always larger than that of a_1. The $se(\hat{a}_2) < se(\hat{a}_1)$ but $se(\hat{a}_2)$ decreased as $\rho(\theta_1,\theta_2)$ increased and $se(\hat{a}_1)$ increased. As $\rho(\theta_1,\theta_2)$ increased, the means of both $\hat{a}_1$ and $\hat{a}_2$ increased and the

standard deviations both increased. There were larger AAD($\hat{a}_1$) than in the A data sets but smaller AAD($\hat{a}_2$). While the statistics for the discrimination parameters were more similar in the A and B data sets than those for the ability or difficulty parameters, they were also more distorted in that the estimates were less like the true values in all cases.

The parameter MDISC was overestimated only in the B3 data set. The mean of this parameter was better recovered as $\rho(\theta_1,\theta_2)$ increased but the standard deviation was increasingly overestimated and was not as well recovered as in the A data sets. Results for the angle recovery were similar to those found in the A data sets. There again seemed to be an attempt to cover the entire space in estimation of parameters related to discrimination.

Correlation coefficients were used (see Table 5 above) to determine adequacy of the parameter recovery. The correlation between $\hat{a}_1$ and $\hat{a}_2$ was greatly reduced but changes in $\rho(\theta_1,\theta_2)$ did not have an affect on this. The correlations between the parameter and its estimate were much lower than in the corresponding A data sets for both a_1 and a_2. The relationship between multidimensional discrimination parameter MDISC and its estimate was also reduced. Differentiated ability on θ_2 did affect recovery of the discrimination parameters.

Part 3

Interaction Effects of Correlated Abilities and a Differentiated Ability

There were four possible interaction effects, the first in the recovery of $\rho(\theta_1,\theta_2)$. There was an interaction between correlation of abilities and differentiated ability on θ_2 on the estimated correlation of abilities. For the B data sets, there was little effect of correlated abilities. For the A data sets, much steeper slopes resulted when $r(\hat{\theta}_1,\hat{\theta}_2)$ was plotted against $\rho(\theta_1,\theta_2)$ (Figure 1). There was a poorer recovery of $\rho(\theta_1,\theta_2)$ in the B data sets with the exception of B2. The

difference in $r(\hat{\theta}_1,\hat{\theta}_2)$ was small between data set A2 and B2 and perhaps is not as meaningful. Indeed this may not be a true interaction even though the lines cross as the B data sets consistently appear to recover $\rho(\theta_1,\theta_2)$ less well. The slightly better recovery of the $\rho(\theta_1,\theta_2)$ of 0.25 may in fact be an artifact of a regression line showing no relationship and consistently estimating correlation close to 0.25 regardless of the true correlation. One would expect $\rho(\theta_1,\theta_2)$ to be better recovered in the full distribution of θ_2 at any level of correlation.

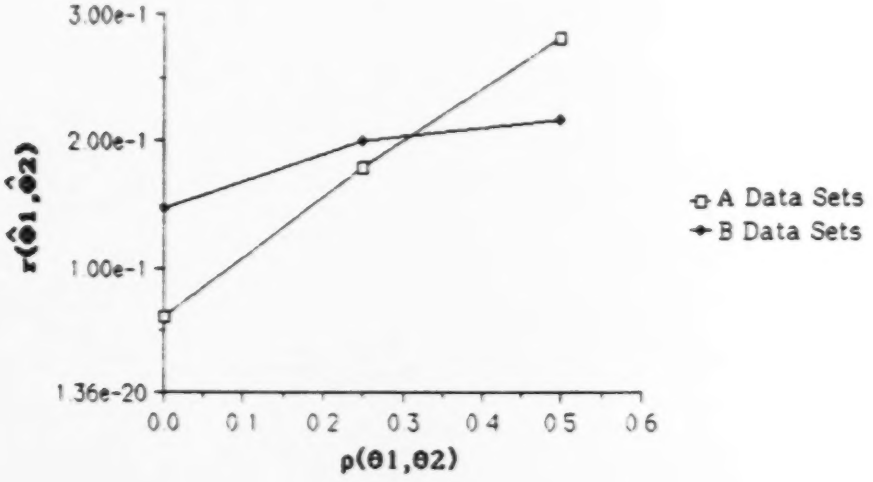

Figure 1. The relationship between $\rho(\theta_1,\theta_2)$ and $r(\hat{\theta}_1,\hat{\theta}_2)$ for the six data sets.

A second interaction occurred between correlation of abilities and differentiated ability on the correlation of each ability estimate with its parameter. As $\rho(\theta_1,\theta_2)$ increased, $r(\theta_1,\hat{\theta}_1)$ decreased while $r(\theta_2,\hat{\theta}_2)$ increased. Increased $\rho(\theta_1,\theta_2)$ had more adverse affects on $r(\theta_1,\hat{\theta}_1)$ than $r(\theta_2,\hat{\theta}_2)$. The $\hat{\theta}_2$ appeared to depend more on θ_1 ability (i.e., $r(\theta_1,\hat{\theta}_2)$ increased as $\rho(\theta_1,\theta_2)$ increased and was larger in the B data sets than in the A data sets). This was not the case with $\hat{\theta}_1$ which did not seem to depend on θ_2 in either A or B data

sets. The $\theta_1\theta_2$-space would appear to be collapsing. The distribution of the discrimination parameters may be contributing to this result as much as differentiated ability and correlation between abilities.

A third interaction was found as $se(\hat{a}_1)$ and $s(\hat{a}_1)$ were affected by correlation of abilities and differentiated ability on θ_2. As $\rho(\theta_1,\theta_2)$ increased, the $s(\hat{a}_1)$ decreased in the A data sets but increased in the B data sets, whereas $s(\hat{a}_2)$ increased in both the A and B data sets. In the A data sets, the $se(\hat{a}_1)$ decreased as $\rho(\theta_1,\theta_2)$ increased but increased in the B data sets. The $se(\hat{a}_2)$ decreased as $\rho(\theta_1,\theta_2)$ increased in both A and B data sets. Increasing the degree of correlation between abilities and a differentiated θ_2 ability combine to give poorer recovery of a_1. While it would be expected that the recovery may deteriorate in B data sets, it was not expected that increasing $\rho(\theta_1,\theta_2)$ would cause further deterioration. As the abilities became more correlated, more information is being used to estimate the second dimension ($se(\hat{a}_2) < se(\hat{a}_1)$ and $se(\hat{a}_2)$ decreases as $\rho(\theta_1,\theta_2)$ increases). As well, the $AAD(\hat{a}_1)$ both increased as correlation increased. These results might be related to the recovery of the mean of a_2 as being larger than the mean of a_1 and to the possible collapsing of the space. Clearly, the B samples didn't cover the ability space adequately. The rescaling of the θ_2 may be contributing to this interaction.

A fourth interaction was found between the correlation of abilities and differentiated ability on θ_2 affecting the mean of MDIF. Surprisingly, in the A data sets, as $\rho(\theta_1,\theta_2)$ increased, $\overline{\hat{MDIF}}$ changed very little. In the B data sets, as $\rho(\theta_1,\theta_2)$ increased, $\overline{\hat{MDIF}}$ decreased (the items appear to be getting easier). This was as expected. Since MDIF is a function of d and MDISC, and a_2 (a part of MDISC) was better estimated in the B data sets, this may explain why $\overline{\hat{MDIF}}$ became smaller (indicating easier items) but $\overline{\hat{d}}$ did not change. A differentiated

ability on θ_2 affected the size of the difficulty means moreso than the degree of correlation.

Conclusions

This research study was designed to determine how well multidimensional IRT ability and item parameters would be estimated under certain specified conditions. The conditions were different degrees of correlation between the two ability dimensions and a differentiated ability on a second dimension.

The results of the research indicated that as the ability dimensions became more correlated, there was a tendency for the two-dimensional ability space to collapse. MIRTE tended to underestimate the degree of correlation between the ability dimensions but did not force orthogonality on the dimensions. Of the item parameters, the difficulty parameter was recovered most successfully. As the ability dimensions became more highly correlated, the discrimination parameter estimate for the predominant dimension (a_1) was underestimated while discrimination on the second dimension (a_2) was overestimated. The discrimination parameters in general were not well recovered. Increasing the correlation between the ability dimensions tended to result in even poorer recovery of the discrimination parameters. For correlated dimensions the effects of item structure and ability structure were compounded as found by McKinley and Reckase (1984). The discrimination parameters did not cover the latent space adequately. In the recovery there was a tendency to spread the discrimination parameters over the entire latent space. This also occurred with the ability estimates and would indicate some rotational indeterminacy in the recovery of the multidimensional correlated latent space.

Restrictions on the second ability dimension resulted in poorer estimation for parameters of both ability dimensions. The differentiated ability on θ_2 appeared to cause a large shift in the estimates of d, underestimating the mean but retaining the internal structure of the item difficulties. The restrictions on the second ability dimension made the recovery of the discrimination parameters much worse than in the A data sets. The rescaling of the θ_2 estimates clearly affected the parameter recovery for the B data sets, particularly item difficulty.

Four interaction effects of correlation of abilities and a differentiated ability on θ_2 were noted. The correlation of abilities and differentiated ability on θ_2 affected recovery of $\rho(\theta_1,\theta_2)$, recovery of the $r(\theta_1,\hat{\theta}_1)$, the discrimination parameters (in $s(\hat{a}_1)$, $se(\hat{a}_1)$, and $AAD(\hat{a}_1)$), and the mean of the estimate of MDIF. The rescaling of $\hat{\theta}_2$ and the poor coverage of the ability space and the item space partially explained these effects.

As for the analogy of the ESL students, would those students be penalized in placement based on the results of this test? Clearly their raw scores on the tests were lower. As the ability dimensions became more correlated, the raw score for these students improved only slightly. McKinley and Reckase (1984) reported that $\rho(\theta_1,\theta_2)$ was an important factor in the latent ability structure. In terms of the recovery of the primary ability dimension, θ_1, the ESL students portrayed in the B data sets would have poorer recovery of this dimension as indicated by $r(\theta_1,\hat{\theta}_1)$ and $AAD(\hat{\theta}_1)$. If the M2PL model were chosen to represent the response data and MIRTE were used to analyze the data, these students would probably be penalized if their θ_1 estimates were used to determine placement. However, because of the rescaling, the question of how the ability estimates of the ESL students are affected cannot really be determined. If the A

and B data sets had been pooled together, it would have mirrored a more realistic educational situation.

There are three issues of concern identified in this research: the problems caused by the rescaling of the θ_2 estimates, the recovery of the two-dimensional space, and the dimensionality of the items.

The rescaling of the θ_2 estimates in the B data sets affected estimates of difficulty as well as estimates of thetas and discriminations. The estimates of means of d and MDIF were adversely affected in the B data sets. However, correlations between the parameters and the corresponding estimates were good. The estimates of the mean of a_2 improved in the B data sets. It cannot be determined from the results reported here the extent of the effects of rescaling but it appears that the rescaling problem affects all parameter estimates somewhat.

The recovery of the structure of the ability space is also a concern. There was a tendency for the space to collapse as the abilities became more correlated. This may relate to a rotational indeterminacy in the recovery of the abilities. In the initial research design, some items pure on the second dimension were included in order to anchor the abilities in an attempt to improve the recovery of all parameters. Since such a test would not simulate the desired condition, this decision was not made. This might be reconsidered in a future design. The collapsing of the space as $\rho(\theta_1,\theta_2)$ increased not only affected the theta estimates but also the discrimination estimates. In the B data sets, the structure of the latent space was recovered less well than in the A data sets. In retrospect, combining corresponding A and B data sets prior to analysis of the raw score vectors would provide a sample which more typically represents the situation in which ESL students would likely be placed and would have allowed for better coverage of the $\theta_1\theta_2$-space. This might improve

the estimation of some parameters and it would also eliminate the rescaling problem.

The third issue is the dimensionality of the item space. Twenty-six of the items were unidimensional (pure on a_1). The remaining 78 were two-dimensional, 52 requiring more ability on the first dimension for a correct response, 26 requiring equal amounts of both abilities. The latent structure of the data was more complex than a two-dimensional test composed of two sets of unidimensional items. There were serious concerns with respect to the recovery of the item space, the most serious being the apparent dominance of a_2 over a_1, or α_2 over α_1. The poor recovery of the discrimination parameters also affected recovery of the difficulty and ability parameters. The item space seemed to become somewhat unidimensional. The estimates of the a_is were more alike and the size of the α_1 angles moved towards 45° with α_2 becoming dominant. Since the range of a_2 was greater than that of a_1, this could have affected the dominance of a_2 over a_1.

Interpretation of parameter estimates appears to depend on the model, $\rho(\theta_1,\theta_2)$, and the characteristics of the data set. There is every indication from the results of this research that there are indeed three components of multidimensionality (subject dimensionality, test dimensionality, and the interaction of the two) as suggested by McKinley and Reckase (1984). Although the population may be multidimensional, if the test is largely unidimensional, resulting scores may tend to unidimensionality as well. It may be expecting too much of the model and MIRTE to have better recovery of the parameters relating to the second dimension when few items measured that dimension and when the populations in the B data sets were low on ability in the second dimension.

Several questions remain at the conclusion of this research which suggest future studies. These are summarized briefly.

Are the results affected by the estimation procedures and/or the model chosen? Replication of the research using different models (perhaps the M3PL model of Bogan and Yen (1983) or a noncompensatory model) would indicate to what extent model choice affected results. Inclusion of a guessing parameter in the model would provide additional information. A more recent version of MIRTE allows for inclusion of the c-parameter

It would be useful as well to estimate item parameters only while holding the given ability parameters fixed and vice versa to determine further the efficiency of the MIRTE program. These results could be compared with those obtained when item and ability parameters are simultaneously estimated. Presumably both item and ability parameters would be better estimated. However, one could study the effects of each by varying the other parameters, i.e., specifying different conditions for item parameters in order to determine the effects on the ability estimates and vice versa.

Corresponding A and B data sets could be combined in order to present the ESL-type group in a large sample of wider variability more typical of a real life situation. This should solve some of the rescaling and space problems.

The test design might be altered to allow for better distribution of the discrimination parameters. The discrimination and difficulty parameters might be randomly generated to cover the space. The test would then not simulate the condition that it primarily measure one of the two dimensions. However, valuable information might be gained on parameter recovery.

It would be useful to determine how well the ability dimensions were recovered at different ability levels rather than just at the mean level of ability although the standard errors, average absolute deviations, and correlations do give some indication of overall recovery. This could be ascertained by looking at the θ-vectors in different sections of the $\theta_1\theta_2$-space and comparing the original

(θ_1,θ_2) with its estimate. It would also be useful to know how influential the second ability dimension became as the items required more of this ability for a correct response.

Another area of interest is that of item difficulty. Further analysis of the examinee results on easy versus difficult items at different ability levels would provide useful information for test builders.

A test with a wider range of discrimination values could determine how discrimination values affect recovery of item and ability parameters. Analysis of discrimination parameter recovery in different areas of the ability space could also be useful. Providing more items requiring both dimensions and some items pure on both dimensions would provide some indication of how the discrimination values need to be chosen to improve estimates. The poor recovery of the discrimination parameters is a cause for concern.

This research study provides encouraging results for those working in multidimensional item response theory. An important finding is the capability of MIRTE to retain the structure of the data and the people. Although there was some tendency to collapse the latent space as $\rho(\theta_1,\theta_2)$ increased, estimates provided by MIRTE recovered two dimensions. It would be judicious to further develop estimation programs so that rotational solutions could be produced which might alleviate the tendency to collapse a two-dimensional space as the correlation between the dimensions increases.

References

Ackerman, T. A. (1985). *M2PLGEN: A computer program for generating thetas and response strings corresponding to the M2PL model*. Iowa City, Iowa: American College Testing.

Ackerman, T. A. (1987, April). *The use of unidimensional item parameter estimates of multidimensional items in adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Ansley, T. N. (1984). An empirical investigation of the effects of applying a unidimensional latent trait model to two-dimensional data. (Doctoral dissertation, University of Iowa, 1984). *Dissertation Abstracts International, 45/07, 2074A*.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9(1)*, 37-48.

Bogan, E. D., & Yen, W. M. (1983, April). *Detecting multidimensionality and examining its effects on vertical equating with the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Carlson, J. E. (1987). *Multidimensional Item Response Theory Estimation: A Computer Program* (ACT Research Report 87-19). Iowa City, IO: American College Testing Program.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22(4)*, 249-262.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional IRT models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199.

Greaud, V. A. (1988, April). *Some effects of applying unidimensional IRT to multidimensional tests*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

International Mathematical and Statistical Libraries. (1979). IMSL Library (7th ed.). Houston, TX: Author.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

McCauley, C. D., & Mendoza, J. (1985). A simulation study of item bias using a two-parameter item response model. *Applied Psychological Measurement, 9(4)*, 389-400.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99.

McKinley, R. L. (1983, April). *A multidimensional extension of the two-parameter logistic latent trait model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec. (ERIC Document Reproduction Service No. ED 228 326)

McKinley, R. L., & Reckase, M. D. (1983a). *An application of a multidimensional extension of the two-parameter logistic latent trait model* (ONR-83-3). (ERIC Document Reproduction Service No. ED 240 168)

McKinley, R. L., & Reckase, M. D. (1983b). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15(3)*, 389-390.

McKinley, R. L., & Reckase, M. D. (1984). *An investigation of the effect of correlated abilities on observed test characteristics* (Research Report). Iowa City, Iowa: American College Testing Program, Test Development Division. (ERIC Document Reproduction Service No. ED 249 249)

Muraki, E., & Englehard, G. (1985). Full-information item factor analysis: applications of EAP scores. *Applied Psychological Measurement, 9(4)*, 417-430.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4*, 207-230.

Reckase, M. D. (1985a). *Models for multidimensional tests and hierarchical structured training materials* (Research Report ONR-85-1). Iowa City, Iowa: American College Testing Program.

Reckase, M. D. (1985b, April). *The difficulty of test items that measure more than one ability*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Reckase, M. D. (1986, April). *The discriminating power of items that measure more than one dimension*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D., Carlson, J. E., Ackerman, T. A., & Spray, J. A. (1986, June). *The interpretation of unidimensional IRT parameters when estimated from multidimensional data*. Paper presented at the annual meeting of the Psychometric Society, Toronto, Ontario.

Wilson, D., Wood, R. L., & Gibbons, R. (1984). *TESTFACT: Test Scoring and item factor analysis* [Computer program]. Moorsville, IN: Scientific Software, Inc.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton, NJ: Educational Testing Service.